TOWARDS ENHANCED DIAGNOSIS OF DISEASES USING STATISTICAL ANALYSIS OF GENOMIC COPY NUMBER DATA

An Undergraduate Research Scholars Thesis

by

ISHA ABBASI[1], RAWAN ABDULGADIR[2], WEAM MAZEN[3], NADIN MOHAMED[4], AND ASRA SAEED[5].

Submitted to the LAUNCH: Undergraduate Research office at
Texas A&M University
in partial fulfillment of requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by
Faculty Research Advisor: Dr. Mohamed Nounou

May 2021

Majors:
Chemical Engineering[1]
Chemical Engineering[2]
Chemical Engineering[3]
Electrical Engineering[4]
Computer Engineering[5]

RESEARCH COMPLIANCE CERTIFICATION

Research activities involving the use of human subjects, vertebrate animals, and/or biohazards must be reviewed and approved by the appropriate Texas A&M University regulatory research committee (i.e., IRB, IACUC, IBC) before the activity can commence. This requirement applies to activities conducted at Texas A&M and to activities conducted at non-Texas A&M facilities or institutions. In both cases, students are responsible for working with the relevant Texas A&M research compliance program to ensure and document that all Texas A&M compliance obligations are met before the study begins.

We, Isha Abbasi[1], Rawan Abdulgadir[2], Weam Mazen[3], Nadin Mohamed[4], and Asra Saeed[5], certify that all research compliance requirements related to this Undergraduate Research Scholars thesis have been addressed with my Research Faculty Advisor prior to the collection of any data used in this final thesis submission.

This project did not require approval from the Texas A&M University Research Compliance & Biosafety office.

TABLE OF CONTENTS

Page

ABSTRACT 1

ACKNOWLEDGEMENTS 3

NOMENCLATURE 4

1. INTRODUCTION 6

2. METHODS 9

    2.1 Filtering 9
    2.2 Fault Detection 14

3. RESULTS 18

    3.1 Filtering 18
    3.2 Fault Detection 21
    3.3 Applications of real genomic copy number data 25

4. CONCLUSION 31

REFERENCES 32

APPENDIX: COPY NUMBER DATA (SW837 AND MPE600) 35

ABSTRACT

Towards Enhanced Diagnosis of Diseases using Statistical Analysis of Genomic Copy Number Data

Isha Abbasi[1], Rawan Abdulgadir[2], Weam Mazen[3], Nadin Mohamed[4], and Asra Saeed[5].
Department of Chemical Engineering[1]
Department of Chemical Engineering[2]
Department of Chemical Engineering[3]
Department of Electrical Engineering[4]
Department of Computer Engineering[5]
Texas A&M University

Research Faculty Advisor: Dr. Mohamed Nounou
Department of Chemical Engineering
Texas A&M University

Genomic copy number data are a rich source of information about the biological systems they are collected from. They can be used for the diagnoses of various diseases by identifying the locations and extent of aberrations in DNA sequences. However, copy number data are often contaminated with measurement noise which drastically affects the quality and usefulness of the data. The objective of this project is to apply some of the statistical filtering and fault detection techniques to improve the accuracy of diagnosis of diseases by enhancing the accuracy of determining the locations of such aberrations. Some of these techniques include multiscale wavelet-based filtering and hypothesis testing based fault detection. The filtering techniques include Mean Filtering (MF), Exponentially Weighted Moving Average (EWMA), Standard Multiscale Filtering (SMF) and Boundary Corrected Translation Invariant filtering (BCTI). The fault detection techniques include the Shewhart chart, EWMA and Generalized Likelihood Ratio

(GLR). The performance of these techniques is illustrated using Monte Carlo simulations and through their application on real copy number data. Based on the Monte Carlo simulations, the non-linear filtering techniques performed better than the linear techniques, with BCTI performing with the least error . At an SNR of 1, BCTI technique had an average mean squared error of 2.34% whereas mean filtering technique had the highest error of 5.24%. As for the fault detection techniques, GLR had the lowest missed detection rate of 1.88% at a fixed false alarm rate of around 4%. At around the same false alarm rate, the Shewhart chart had the highest missed detection of 67.4%. Furthermore, these techniques were applied on real genomic copy number data sets. These included data from breast cancer cell lines (MPE600) and colorectal cancer cell lines (SW837).

ACKNOWLEDGEMENTS

Contributors

We would like to thank our mentors, Dr. Mohamed Nounou, and Dr. Mohammed Ziyan Sheriff, for their guidance and support throughout the course of this research.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

Finally, thanks to our friends for their encouragement and to our parents for their patience and love.

The MATLAB code used and modified for Towards Enhanced Diagnosis of Diseases using Statistical Analysis of Genomic Copy Number Data were provided by Dr. Mohammed Ziyan Sheriff and Dr. Mohamed Nounou.

All other work conducted for the thesis was completed by the students independently.

Funding Sources

No funding was requested or required to conduct this research.

NOMENCLATURE

b_i	A sequence of weighing coefficients
I	Filter length
J	Maximum decomposition length
j	Discretized dilation parameter
k	Translation parameter
L	Width of the control limits
m	Size of moving window
n	Length of the signal
x	Measured data point
$\bar{x}$	Sample mean
$\hat{x}$	Mean filtered data point
Z	Likelihood ratio test statistic
z	Smoothing parameter
α	Smoothing parameter
σ	Standard deviation
λ	Smoothing parameter
μ	Mean
μ_0	Mean of null hypothesis
$\hat{\mu}_{i,\tau,k}$	Maximum likelihood estimates of mean of alternative hypothesis
τ	Observation number that provides the maximum GLR statistic
ϕ	Orthonormal scaling function

ψ	Orthonormal wavelet functions

1. INTRODUCTION

Genomic instabilities, such as the amplification or deletion of chromosomal segments, are often associated with the development of various diseases. For example, in cancer, deletions may influence inactivation of tumor suppressor genes, while amplifications may influence the activation of oncogenes in genomes. Both deletions can cause changes in copy numbers of the tumor DNA. Therefore, proper diagnosis of diseases requires accurate detection of the presence and location of aberrations in DNA sequences. However, dealing with copy number data is not a simple task since they are often riddled with measurement noise which drastically affects the quality and usefulness of the data [1]. Moreover, DNA copy numbers at adjacent probes along the length of a chromosome may often exhibit spatial dependence. This is because the copy number gain or loss at one particularly probe location increases the likelihood of gain or loss at adjacent probe locations.

Microarray-based methods and advances in DNA sequencing technology have created more opportunities to detect CNVs accurately [2], [3]. However, these methods still have some limitations, and the complexity of data samples adds to the challenge. For instance, the complexity of tumor samples has made the detection of cancer specific CNVs even more difficult. These limitations indicate a need for developing more efficient and precise CNV detection methods that employ appropriate normalization and de-noising techniques [4].

Improved detection of aberrations in copy number data can be achieved using various data analysis approaches, such as univariate filtering and fault detection. Various filtering techniques are dependent on existing process models or empirical models. However, since accurate models are not always readily available, especially for biological systems, several

model-free filtering methods have been developed. These filters depend on the information regarding the nature of the errors and the smoothness of the signal. Examples of this are low pass filters which include the Finite Impulse Response (FIR) and the Infinite Impulse Response (IIR) filters. The Mean Filter (MF) and the Exponentially Weighted Moving Average (EWMA) [5], [6] are types of FIR and IIR filters respectively. More advanced methods that rely on wavelet-based multiscale representation of data have also been developed and used to analyze genomic data [7]. Multiscale based filtering provide advantages as multiscale representation of data allows efficient separation of deterministic and stochastic features in data. In this work, these methods will be utilized to enhance filtering different copy number data sets representing various diseases.

Furthermore, a number of univariate fault detection techniques, along with their multivariate techniques have been developed to monitor and detect faults for various applications. These techniques include the Shewhart technique, Cumulative Sum (CUSUM), Exponentially Weighted Moving Average "EWMA" [8]–[10]. More recently, statistical hypothesis testing techniques such as the generalized likelihood ratio (GLR) have been utilized to enhance fault detection, as they are able to utilize available data to monitor faults by computing maximum likelihood estimates [11]–[13]. The GLR technique is able to detect changes in the mean and/or variance, depending on the requirement [14]. Multiscale wavelet-based representation of data has often been utilized as they are able to denoise data efficient, and provide a number of additional advantages [7], [15]. In previous efforts, some of these techniques have been applied to detect aberrations in copy number data, which include multiscale Shewhart chart [16]. In this study, these methods are applied on average log2 ratios of copy number data. Generally, zero mean refers to a healthy sequence. Instances where the mean

is greater or lower than zero are thus associated with aberrations that can result in diseases. These methods help identify positions along the genome where the aberration exist and therefore help in disease diagnosis. In this work, the advantages of multiscale representations and a hypothesis based technique will be utilized to enhance the detection of aberrations in copy number data. The advantages of these techniques will be illustrated through their application using various copy number data sets.

2. METHODS

2.1 Filtering

2.1.1 Mean Filtering (MF)

Mean filtering is a type of linear filtering technique that filters the signal by computing the weighted sum of previous measurements in a window of finite length. Therefore, it is a finite impulse response (FIR) filter with a finite window size and finite impulse response [7]. Mean filter is computational efficient and easy to implement, making it a popular filtering technique. Moreover, linear filters are low pass filters with a selected cutoff frequency and can be expressed as,

$$x_t = \sum_{i=0}^{I-1} b_i x_{t-i} \tag{2.1}$$

where I is the filter length, b_i is a sequence of weighing coefficients that satisfies the condition $\sum_i b_i = 1$. Mean filters require all weighing coefficients b_i to be equal, $b_i = \frac{1}{I}$. Therefore, for a mean filter length of I, a mean filtered data point is represented by the average of the last I data points as shown below[17].

$$x_t = \frac{1}{I}(x_t + x_{t-1} + \cdots + x_{t-I+1}) \tag{2.2}$$

For the case studies, the optimum mean filter length is estimated using cross validation by testing different mean filter lengths and using the one with the smallest mean square error.

2.1.2 Exponentially weighted moving average (EWMA)

Exponentially weighed moving average (EWMA) is an Infinite Impulse Response (IIR) filter. IIR is a low pass, model-free, linear filter that has an infinite filter length [17]. EWMA

filter smoothes a data point by exponentially averaging that data point with all previous measurements. As it is a low pass filter it removes high frequency components in the measured signal. Computationally, it is implemented using,

$$\hat{x}_t = \alpha x_t + (1-\alpha)\hat{x}_{t-1} \tag{2.3}$$

where the parameter α is a smoothing parameter lying between zero and unity. A value of zero corresponds to keeping only the first measured data point while a value of one indicates no smoothing. Equation (2.3) can also be represented as,

$$\hat{x}_t = \sum_{i=0}^{\infty} \alpha(1-\alpha)^i \hat{x}_{t-i} \tag{2.4}$$

The filter coefficients drop exponentially depending on α giving more significance to the recent measurements.

2.1.3 Multiscale filtering (MF)

In multiscale decomposition, a signal is represented at multiple resolutions by decomposing it on orthonormal scaling and wavelet functions. The decomposition of the signal produces a scaled signal and a detail signal at every level. This method of wavelet-based decomposition involves low pass and high pass filters which are applied to the signal to form the first scaled signal and the first detail signal, respectively as seen in **Figure 2-1**. A set of orthonormal scaling functions in the equation below represent the low pass filter,

$$\phi_{ij}(t) = \sqrt{2^{-j}}\,\phi(2^{-j}t - k), \tag{2.5}$$

where j is the discretized dilation parameter and k is the translation parameter [18].

However, the detail signal is projected onto a set of orthonormal wavelet functions represented by the equation below which also represents the high pass filter,

$$\psi_{ij}(t) = \sqrt{2^{-j}}\,\psi(2^{-j}t - k). \tag{2.6}$$

The original signal can be retained by summing all the detail signals and the last scaled signal, and this is represented by the expression,

$$x(t) = \sum_{k=1}^{n2^{-J}} a_{Jk}\phi_{Jk}(t) + \sum_{j=1}^{J}\sum_{k=1}^{n2^{-j}} d_{jk}\psi_{jk}(t), \tag{2.7}$$

where J is the maximum decomposition depth and n is the length of the signal x(t) [17], [18].

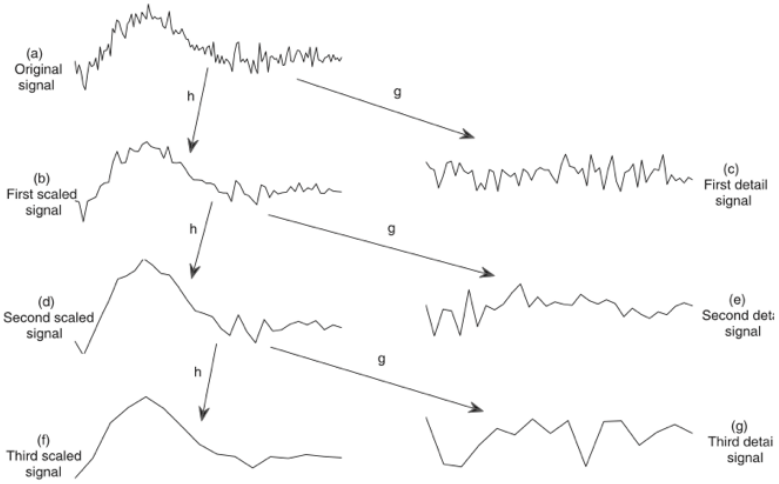

Figure 2-1: A schematic diagram of data decomposed into scaled and detail signals [17]

Multiscale wavelet decomposition involves thresholding the coefficients which can be done by soft or hard thresholding. This is to eliminate the stationary Gaussian noise present in the noisy signal [17]. Soft thresholding shrinks the coefficient values towards the threshold value by subtracting from them, in contrast to hard thresholding which keeps all the coefficient values that are outside the bands of the threshold window and sets the remaining coefficients to zero. Soft thresholding is the method of thresholding being utilized in the multiscale decomposition MATLAB code. Moreover, the function *wavedec* is utilized to perform a 1-D wavelet decomposition for the input signal at a specified optimum depth, using the wavelet *wname*. The

optimum decomposition depth is the optimum depth level for the detail signals; this is determined by applying multiscale filtering at different levels and calculating the mean square error for each depth level. The wavelet used in our case studies is *haar* which is a simple, discontinuous wavelet that resembles a step function.

Multiscale filtering allows for effective elimination of noise from useful features by easily cancelling out the unimportant coefficients which are usually the small wavelet coefficients in the detail signals. The ability of the multiscale representation of data to decorrelate autocorrelated data at multiple scales is another important advantage for this method. However, one limitation to multiscaling is that it requires that the data size or the original signal to be of dyadic length (2^n) [17], [18].

2.1.4 Boundary corrected translation invariant (BCTI)

The boundary corrected translation invariant (BCTI) filtering is another multiscale filtering method. TI filtering involves shifting the signal several times, filtering it, and then taking the average of the translations to improve the smoothness of the filtered data. The disadvantage of TI filtering is that it assumes the signal to be cyclic, creating end effects when boundary corrected wavelets are used, as seen in **Figure 2-2** [7]. To overcome this, BCTI filtering is used instead. Another advantage of BCTI filtering is that less data points are averaged [19].

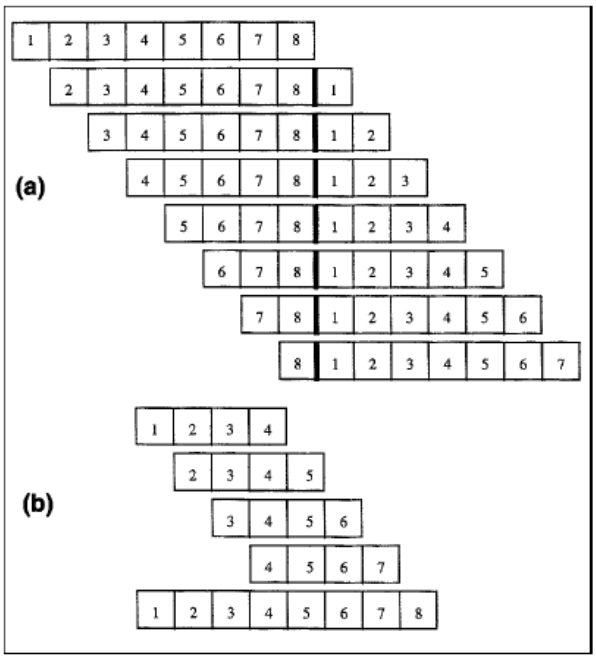

Figure 2-2: Translation mechanisms using in TI (a) and BCTI (b) [7]

Figure 2-3 shows all the filtering techniques applied on the noisy data, for purposes of comparison the noise-free data was also plotted.

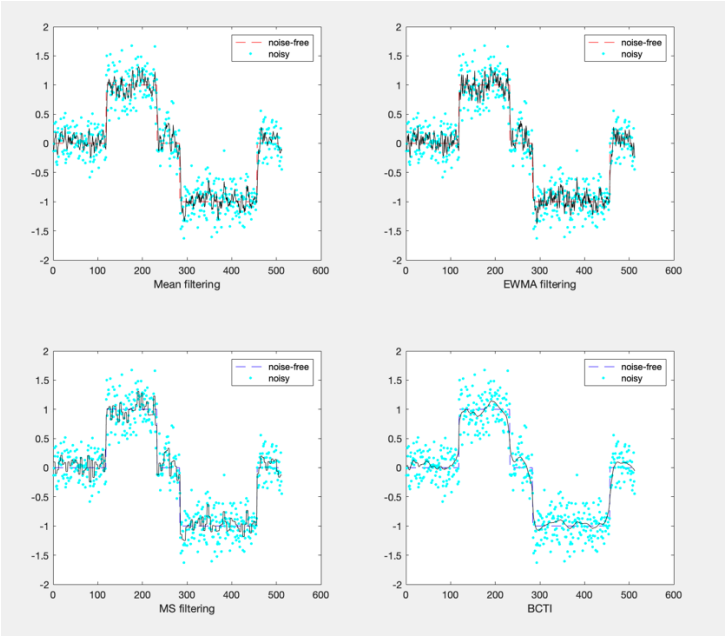

Figure 2-3: Filtering techniques applied to noisy data

As it can be seen in **Figure 2-3,** BCTI was the most successful at filtering data, as it had the most overlap with the noise-free data. Overall, the non-linear techniques, multiscale filtering and BCTI, were more successful at filtering data than the linear techniques, mean filtering and EWMA.

2.2 Fault Detection

2.2.1 Shewhart Chart

The Shewhart chart, developed by Walter Shewhart, is one of the most popular statistical quality control charts because of its simplicity. The Shewhart chart was designed based on the assumptions that the residuals are independent and that the fault-free residuals are normally distributed.

Shewhart charts have three distinct features: Center Line (C), which is typically the mean, Upper Control Limit (UCL), and Lower Control Limit (LCL), which are calculated as follows:

$$UCL = \bar{x} + L\sigma \tag{2.8}$$

$$LCL = \bar{x} - L\sigma \tag{2.9}$$

where $\bar{x}$ is the sample mean, L is the width of the control limits, and σ is the standard deviation of the fault-free residuals. The width of the control limits is usually selected to be 3 for a set of normally distributed fault-free data in order to account for nearly 99.73% of all the deviation.

The main disadvantage of using the Shewhart chart is that it is not very sensitive to change since it does not have memory and deals with every sample independently. This leads to the Shewhart chart only being able to detect faults that are three times the standard deviation [16].

2.2.2 Exponential Weighted Moving Average chart (EWMA)

The conventional exponentially weighted moving average (EWMA) method is a data based, univariate fault detection technique. EWMA utilizes linear filters and applies them on the residuals to improve their sensitivities to small shifts. Moreover, EWMA is less sensitive to the normality assumption as the EWMA statistic is the weighted average of all past and current observations, where the weights assigned to the past observations decrease exponentially. EWMA control scheme involves computation of the EWMA statistic and the upper and lower control limits. EWMA statistic is computed as follows:

$$z_t = f(x_t) = \lambda x_t + (1 - \lambda) z_{t-1} \tag{2.10}$$

where λ is the smoothing parameter which alters the memory of the detection statistic. Likewise, the upper and lower control limits are expressed in terms of the standard deviation of the EWMA statistic and computed as follows:

$$UCL, LCL = \bar{x} + L\sigma \sqrt{\frac{\lambda}{2 - \lambda}} \tag{2.11}$$

where L is the control width, $\bar{x}$ is the sample mean. The choice of the smoothing parameter is made carefully depending on the size of mean shift to be detected. For a larger mean shift, a larger λ is chosen and a smaller λ is used to detect a smaller mean shift. When λ is chosen to be 1, the EWMA statistics only uses the most recent observation. Optimum values for EWMA parameters, L and λ, are chosen based on the size of fault to be detected. In this paper, the parameters are chosen by minimizing ARL_1 and assuming $ARL_0 = 500$, according to the graphs below, obtained from [18].

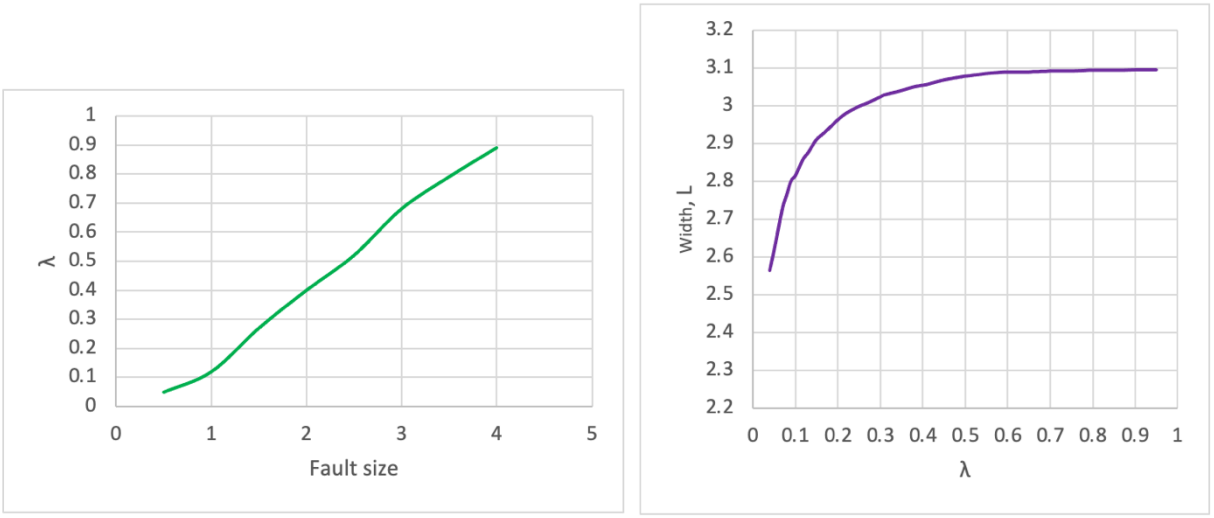

Figure 2-4 EWMA Parameters

2.2.3 Generalized Likelihood Ratio Test (GLRT)

The Generalized Likelihood Ratio Test (GLRT) is a hypothesis testing method meaning it decides which of two hypotheses (null or alternative) best describe a given data set. The application of the GLRT in fault detection first involves the generation of a model which describes the normal operation of signals in a given data set [20]. The null and alternate hypotheses are then defined for the given situation [21].

The GLRT is based on the classical likelihood ratio statistic wherein distribution functions for a given parameters, such as the mean and covariance, corresponding to each hypothesis are assumed. The ratio of the alternative distribution to the null distribution provides the maximum detection probability at a fixed alarm rate and represents the likelihood ratio test statistic, Z [22].

$$Z = \frac{P_1(\theta_1)}{P_0(\theta_0)} = \frac{N(\mu_1, \sigma^2)}{N(\mu_0, \sigma^2)} \qquad (2.12)$$

In the GLRT, the parameters are unknown and are thus replaced by their maximum likelihood estimates, $\hat{\mu}_{i,\tau,k}$. The maximum is taken for a fixed number of samples of a moving window of size *m* that the user specifies. When both the null and alternative hypotheses have a Gaussian distribution, the log-likelihood ratio is simplified and maximized, giving the GLRT statistic.

$$GLR_k = \max_{0 \le \tau < k} \frac{k-\tau}{2\sigma_0^2}\left(\hat{\mu}_{1,\tau,k} - \mu_0\right)^2 \qquad (2.13)$$

Figure 2-5 shows all the fault detection techniques applied on the faulty data.

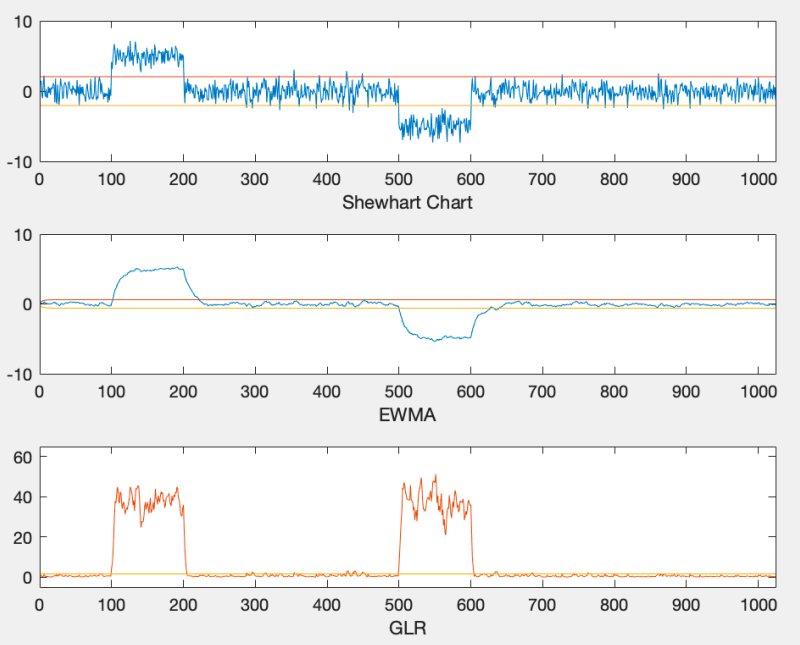

Figure 2-5: Fault detection technique applied to faulty data

3. RESULTS

3.1 Filtering

3.1.1 Monte Carlo Simulation

To compare the performance of the filtering techniques, a Monte Carlo simulation was conducted where generated data having zero mean with two aberrations were used. The simulated data which are assumed to be noise-free are then contaminated with zero mean Gaussian noise. A sample run displaying how the various techniques perform for a signal-to-noise ratio of one can be seen in **Figure 2-3**. By examining **Figure 2-3** it can be seen that BCTI filters and smoothes the signal the most. To further analyze the different techniques, the mean squared error (MSE) and median was calculated. This was done using different signal-to-noise ratios (SNR) varying from 1 to 5. The averages of the mean squared error and the median for 5000 Monte Carlo runs are summarized for each filtering technique in **Table 3-1. Figure 3-1** and **Figure 3-2** summarize the results graphically.

Table 3-1: Comparison of the filtering techniques using a Monte Carlo simulation.

SNR	Technique	Average MSE	Median MSE
1	Mean Filtering	5.2400	5.2000
	EWMA	4.1700	4.1300
	Multiscale Filtering	3.3500	3.2400
	BCTI	2.3400	2.3200
2	Mean Filtering	4.5900	4.5200
	EWMA	3.9600	3.9400

	Multiscale Filtering	3.3200	3.1600
	BCTI	1.7900	1.7700
3	Mean Filtering	0.0300	0.0300
	EWMA	0.0200	0.0200
	Multiscale Filtering	0.0200	0.0200
	BCTI	0.0100	0.0100
4	Mean Filtering	0.0152	0.0151
	EWMA	0.0129	0.0129
	Multiscale Filtering	0.0090	0.0088975
	BCTI	0.0048	0.0048
5	Mean Filtering	0.0198	0.0196
	EWMA	0.0168	0.0167
	Multiscale Filtering	0.0107	0.0108
	BCTI	0.0076	0.0168

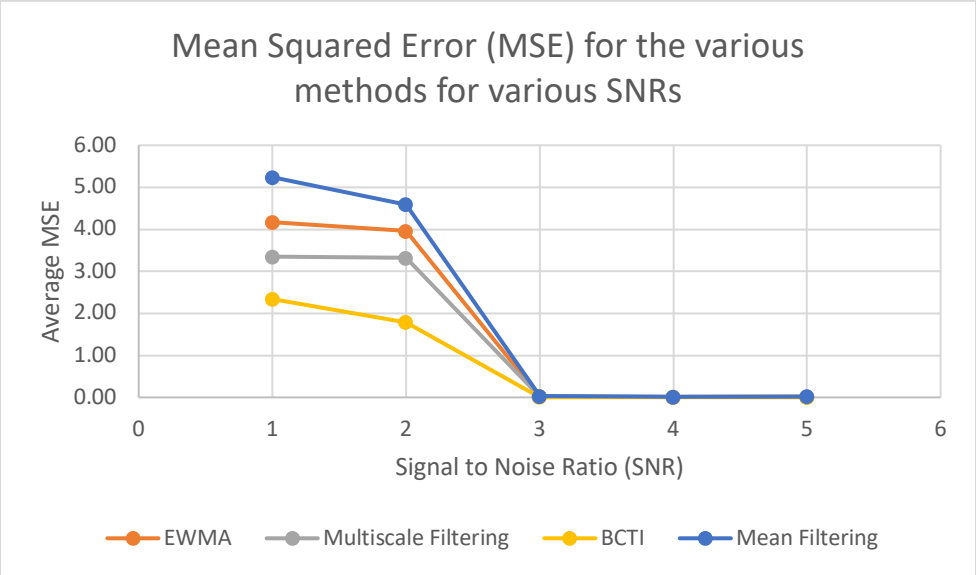

Figure 3-1: Mean Squared Error (MSE) for the various methods at different signal-to-noise-ratios

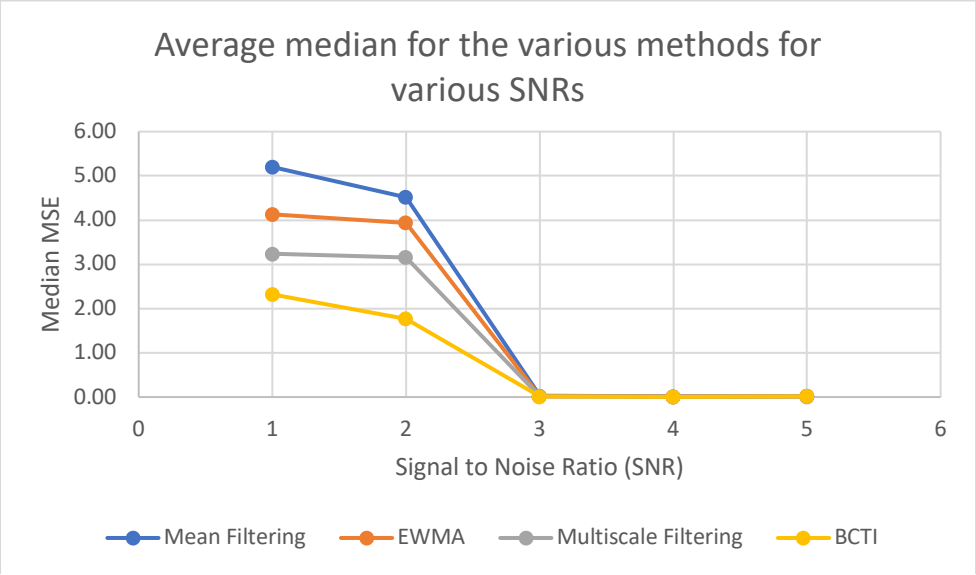

Figure 3-2: Average median for the various methods at different signal-to-noise-ratios

Based on the Monte Carlo simulation, the non-linear methods are more accurate than the linear methods for all the signal-to-noise ratios. This difference drops drastically when SNR is 3 or above as seen in Figure 3-1 and **Figure 3-2,** showing that all the filtering techniques perform well. The MSE and median values for all the filtering techniques were very small and similar for

high SNR as the signal was significantly higher than the noise making filtering easier. The boundary corrected translation invariant (BCTI) technique had the least error whilst the mean filtering method (MF) had the most. As seen in **Figure 3-1,** BCTI consistently had the least mean squared error for every signal-to-noise ratio while mean filtering technique had the highest mean squared error further proving that BCTI is the most accurate filtering technique. Linear techniques performed poorly compared to non-linear filters as they are low pass filters. These filters define a frequency threshold above which all features are considered noise. This creates two issues, first, important features can be deleted due to their high frequency and second, noise can be retained due to their low frequency.

3.2 Fault Detection

3.2.1 Monte Carlo Simulation

To assess the performance of the fault detection techniques, the missed detection (MD) rate, the false alarm (FA) rate, and the out-of-control average run length (ARL_1) were analyzed by performing a Monte Carlo simulation. The missed detection rate refers to the rate at which a fault goes undetected in the faulty region and the false alarm rate is when an observation in the non-faulty region is flagged as a fault. ARL_1 is the number of observations it takes for the fault detection technique to flag a fault in the faulty region is used as a measure of the speed of detection [23]. The Monte Carlo simulation generated data with two aberrations. A sample run displaying how the various techniques perform for a signal-to-noise ratio of one can be seen in **Figure 3-3.**

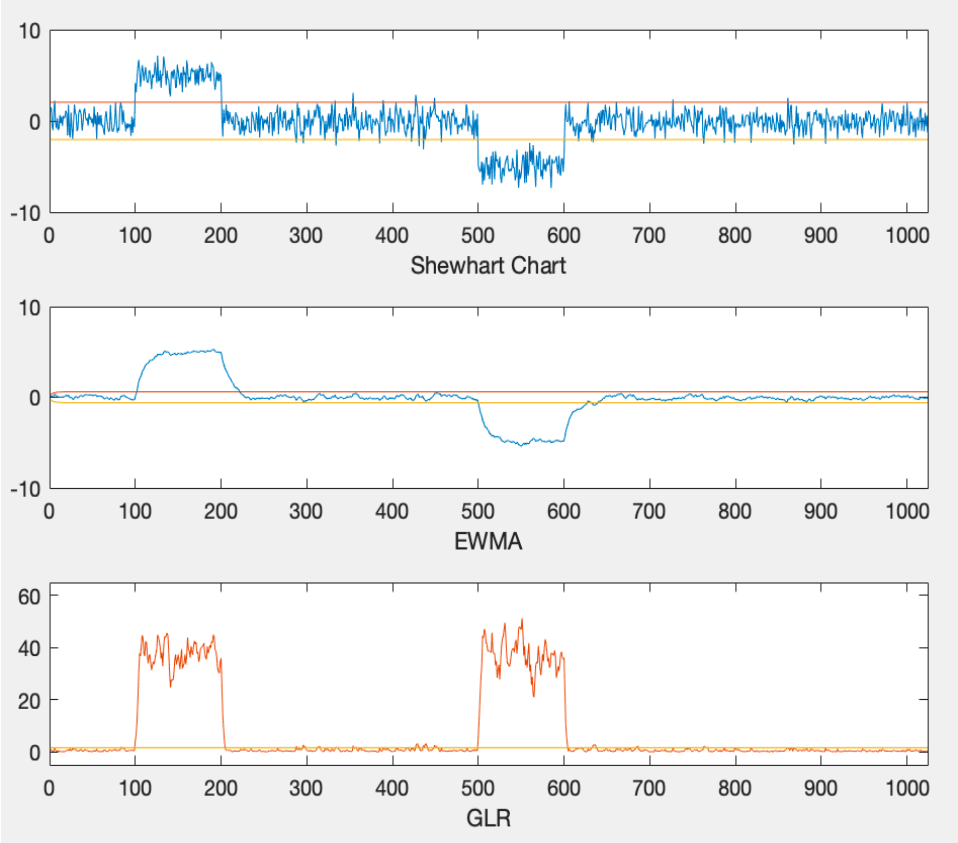

Figure 3-3 Sample run from Monte Carlo simulation for different fault detection techniques

To further analyze each technique, the average of the missed detection (MD) rate, the false alarm (FA) rate, and the out-of-control average run length (ARL_1) for 5000 runs were taken for each fault detection technique. This was repeated for different signal-to-noise ratios ranging from 1 to 5. The results are summarized in **Table 3-2.** Figure 3-1 and **Figure 3-2** summarize the results graphically and display the average missed detection (MD) and the out-of-control average run length (ARL1) for a fixed false alarm rate.

Table 3-2: Comparison of the fault detection techniques using a Monte Carlo simulation.

SNR	**Technique**	**Average FA (%)**	**Average MD (%)**	**Average ARL1 (%)**
1	Shewhart Chart	4.09	67.4	3.20
	EWMA	4.02	2.45	5.74
	GLR	4.35	1.88	4.53
2	Shewhart Chart	4.15	41.7	1.75
	EWMA	5.18	1.45	3.90
	GLR	5.18	0.840	2.66
3	Shewhart Chart	4.13	23.8	1.31
	EWMA	5.83	1.08	3.17
	GLR	5.54	0.49	1.98
4	Shewhart Chart	6.27	9.43	1.11
	EWMA	6.31	0.890	2.77
	GLR	5.78	0.320	1.63
5	Shewhart Chart	5.87	4.87	1.06
	EWMA	5.13	0.68	2.36
	GLR	5.94	0.21	1.43

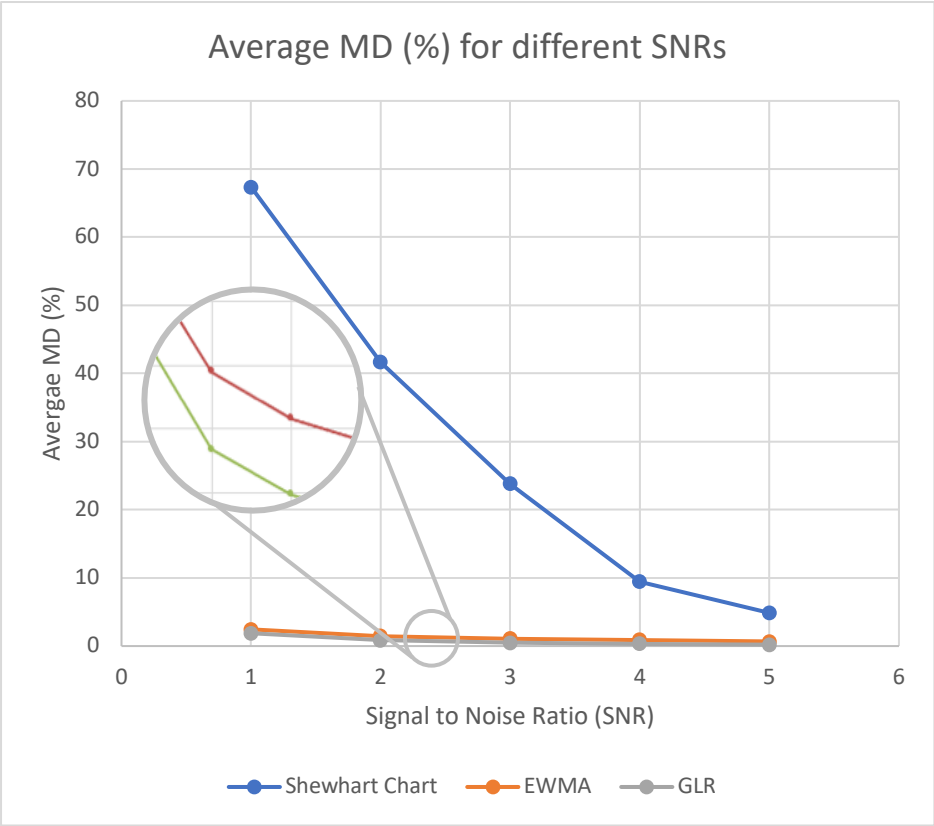

Figure 3-4: Missed detection rate (MD) for the various methods at different signal-to-noise-ratios

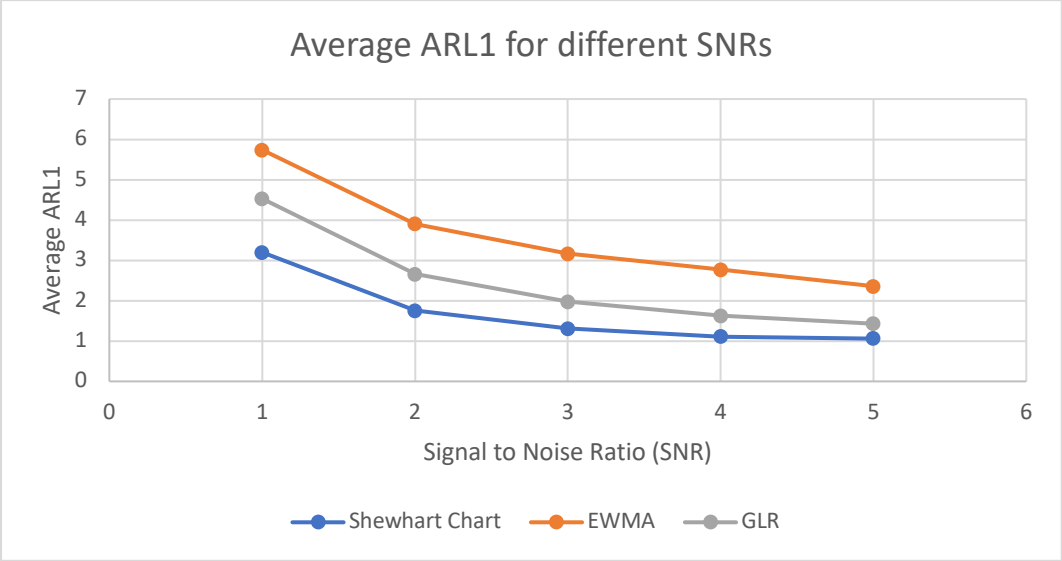

Figure 3-5: Out-of-control average run length (ARL1) for the various methods at different signal-to-noise-ratios

To compare the different fault detection techniques, the false alarm rate was set to be similar across each technique. This was achieved by varying the parameters of each technique. These parameters were the confidence interval for Shewhart Chart, window size and confidence interval for GLR, and lastly L and λ for EWMA. This was done since a trade-off exists between false alarm rate and missed detection rate.

When it comes to the missed detection rate (**Figure 3-4),** as expected, GLR continuously performed the best since it's a hypothesis testing method with its highest missed detection rate being about 2% when the SNR was equal to 1. GLR was followed by EWMA with a maximum of 2.5% and the Shewhart Chart with a maximum of about 67%. As the SNR increases, the difference between the techniques decreased. This is due to the fact that there is less noise and therefore, the faults would be easier to detect.

As for the average run length (**Figure 3-5),** the performance of all the techniques was comparable with the maximum being 3.2 for the Shewhart Chart, 5.7 for EWMA, and 4.5 for GLR. Similar to the missed detection rate, these values had smaller differences as the SNR increased. Overall, the Shewhart Chart was the fastest, followed by GLR then EWMA. This is because the Shewhart Chart processes the data differently to EWMA and GLR since it has no memory. However, the maximum window size of GLR can be increased to decrease the average run length but for the sake of comparison, the false alarm rate was kept constant.

3.3 Applications of real genomic copy number data

Having validated the usefulness of the methods through the extensive Monte Carlo simulations, they were applied on real genomic copy number data. The data set used was log2 ratio data of chromosome 1 obtained from breast cancer cell lines (MPE600) and colorectal cancer cell lines (SW837). This data was plotted against the position in the genome [24]. Since

some of the methods are only applicable on dyadic datasets, some data points with log2 ratio of zero were neglected

3.3.1 Filtering

An immediate observation of the graphs shown in **Figure 3-6** and **Figure 3-7** depicts that the multiscale methods were capable of achieving smoother filtering, with the BCTI filter achieving the best results. The results from the low pass filters on the other hand were rougher and still demonstrated the presence of noise.

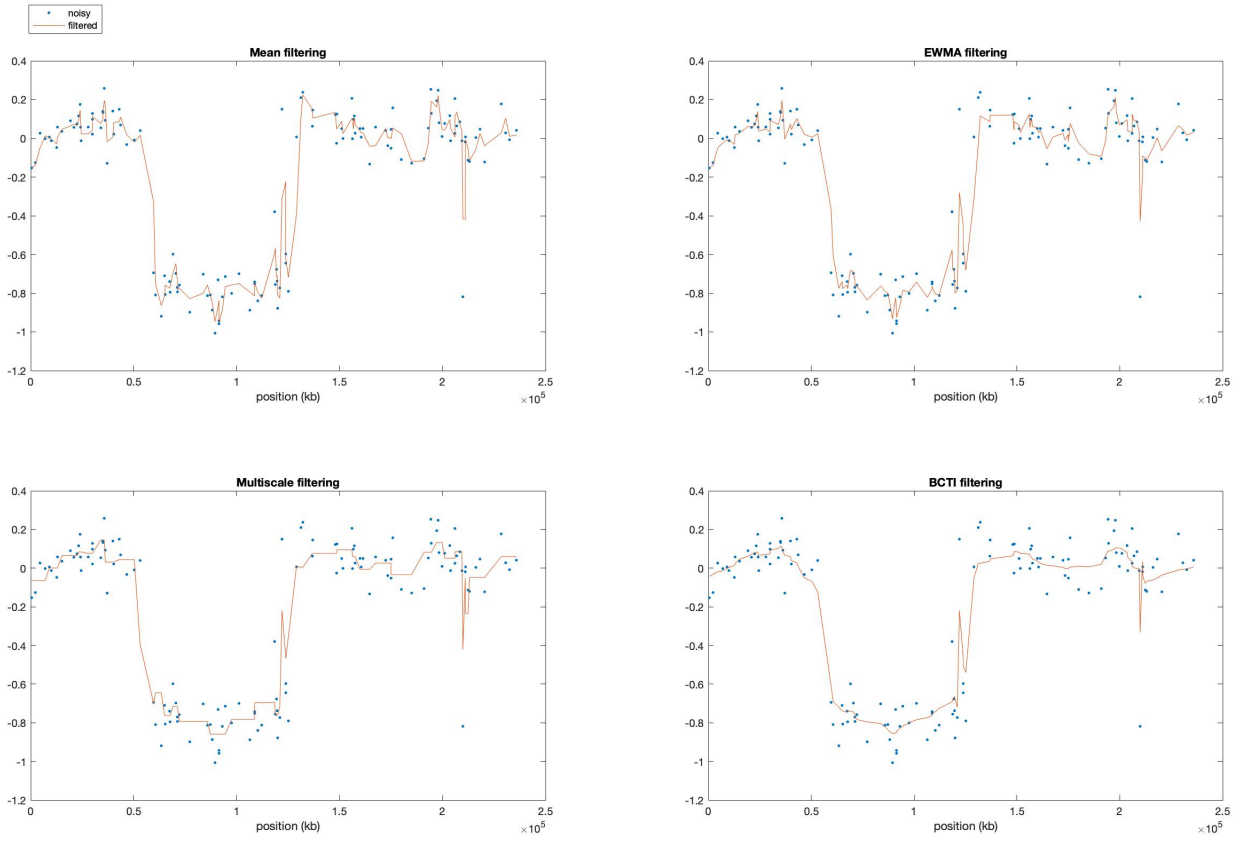

Figure 3-6: Applications of filtering techniques on real genomic copy number data (SW837)

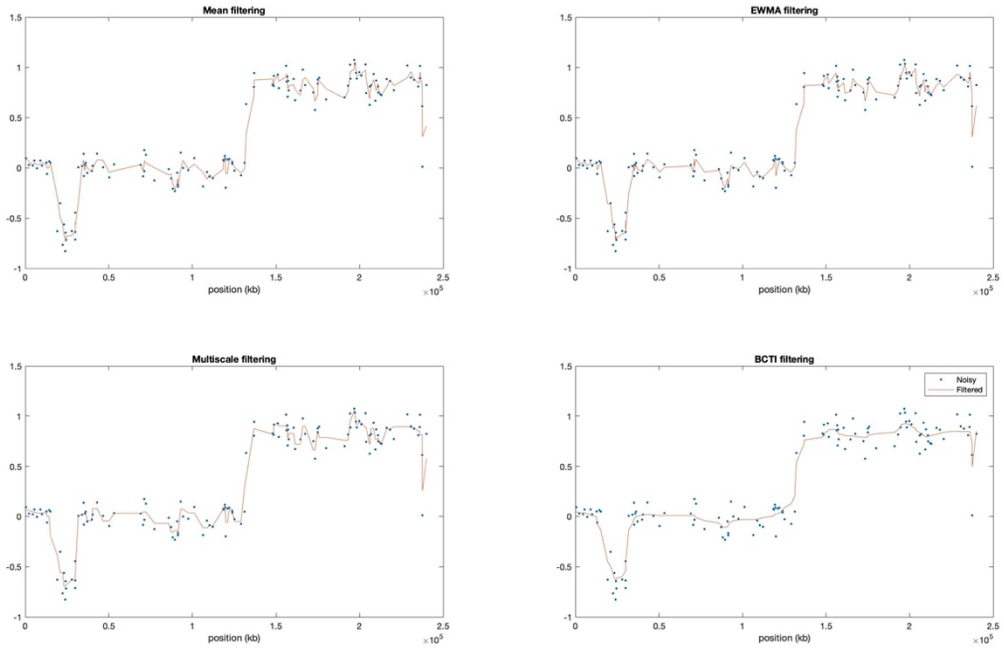

Figure 3-7: Applications of filtering techniques on real genomic copy number data (MPE600)

3.3.2 Fault detection

The application of the fault detection on real data provided the anticipated trend, with the GLR technique outperforming the other techniques. As shown in **Figure 3-8,** the results from the GLR technique clearly depict where the faults are present with no false alarms or missed detections. On the other hand, the graph obtained from the Shewhart method depicts one instance of missed detection relative to the GLR results. Furthermore, it had a higher threshold meaning it provided for a less accurate and precise analysis of the data. Finally, while the EWMA did perform better than the Shewhart method, it did demonstrate one instance of false alarm relative to the GLR results.

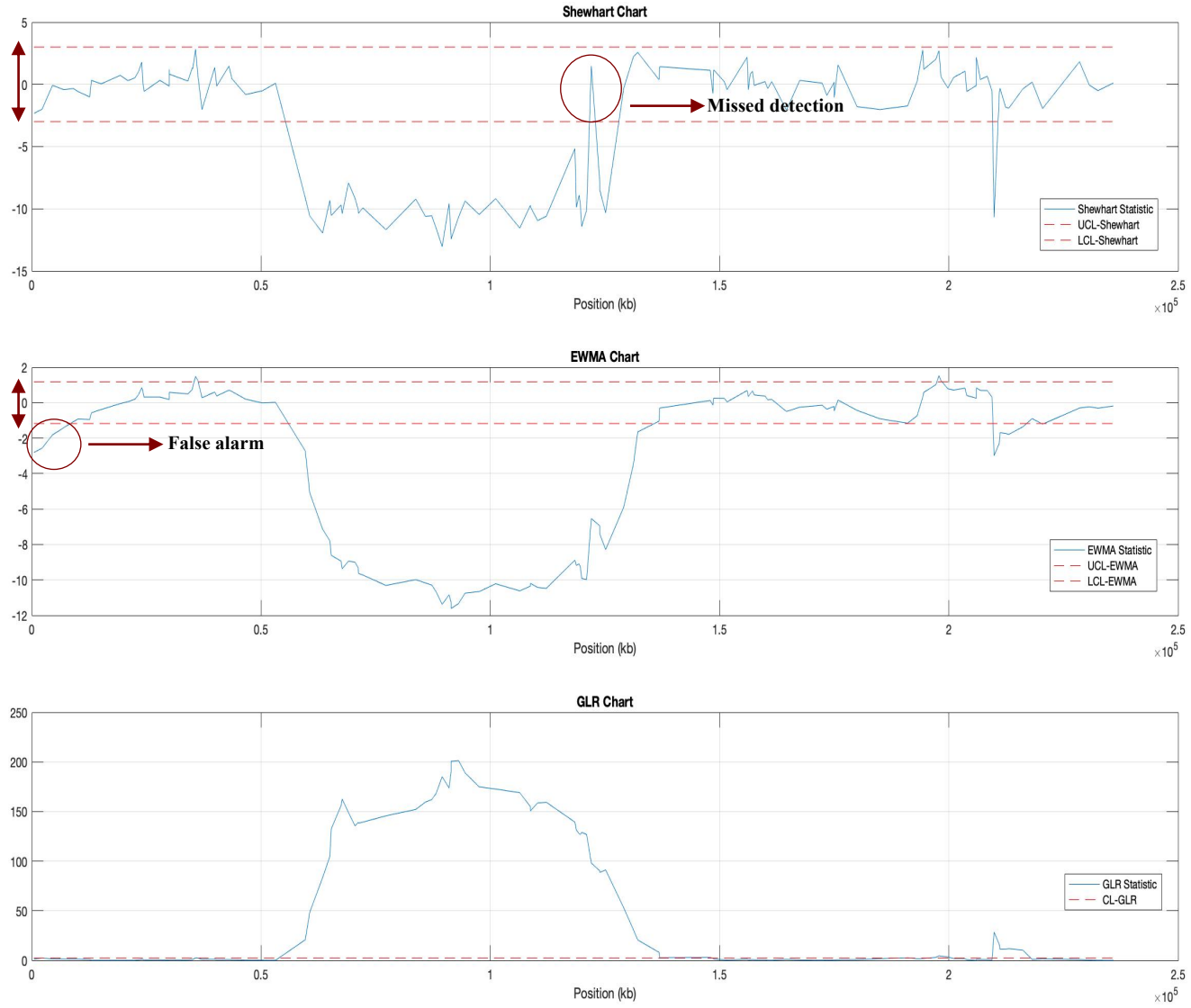

Figure 3-8: Applications of fault detection techniques on real genomic copy number data (SW837)

The fault detection techniques were applied on the breast cancer cell lines and the results can be seen in **Figure 3-9.** These results also depict the previously discussed trend which confirms the effectiveness of these techniques. However, in this case the Shewhart chart performed worse as it depicted more instances of missed detection relative to the GLR results. Furthermore, it had significantly larger threshold limits.

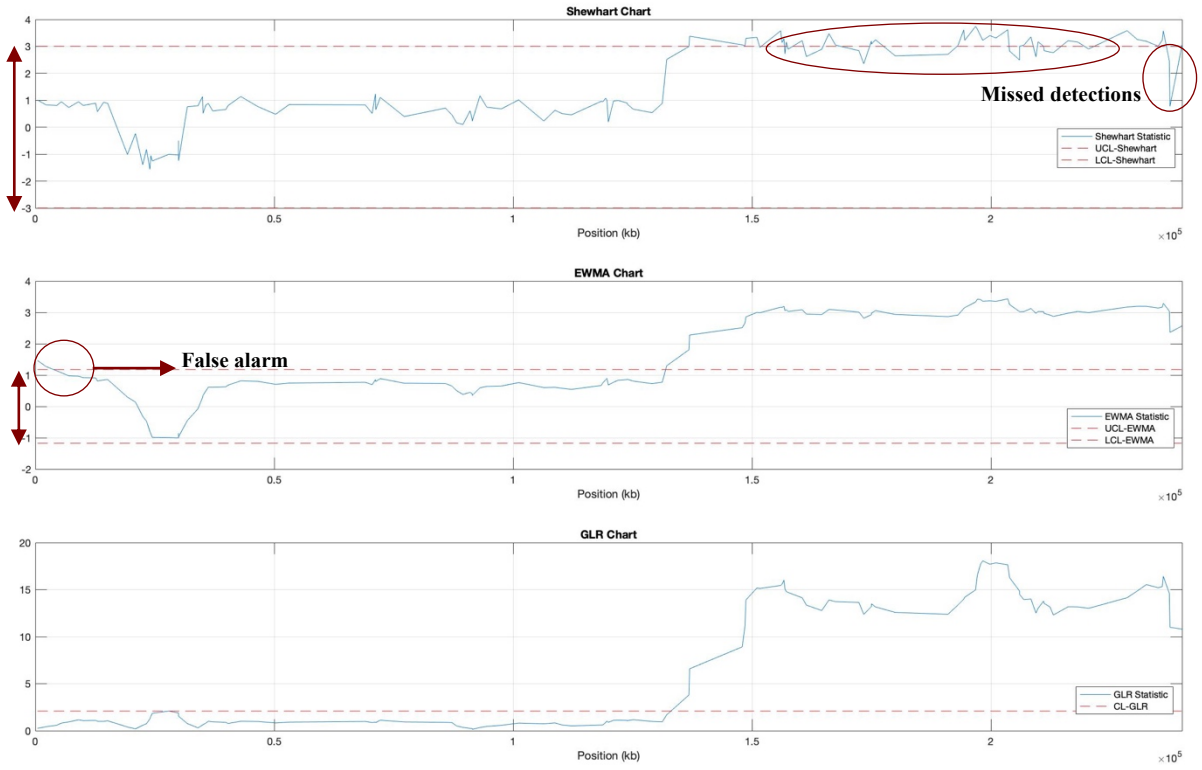

Figure 3-9: Applications of fault detection techniques on real genomic copy number data (MPE600)

3.3.3 Application of fault detection on filtered copy number data

To achieve better results, the use of fault detection methods can be combined with the filtering techniques. To demonstrate this, the Shewhart method was applied on the raw data alone and on the BCTI filtered data. The results shown in **Figure 3-10** and **Figure 3-11** clearly depict that improved results were obtained even with the Shewhart chart which was previously the least effective method. The graph is significantly smoother and accurately depicts where the fault is present. Furthermore, the previously missed fault was detected, and the threshold limits were narrowed down. Overall, the results strongly indicate that these methods can truly be beneficial in detecting aberrations in genomic data at the exact position.

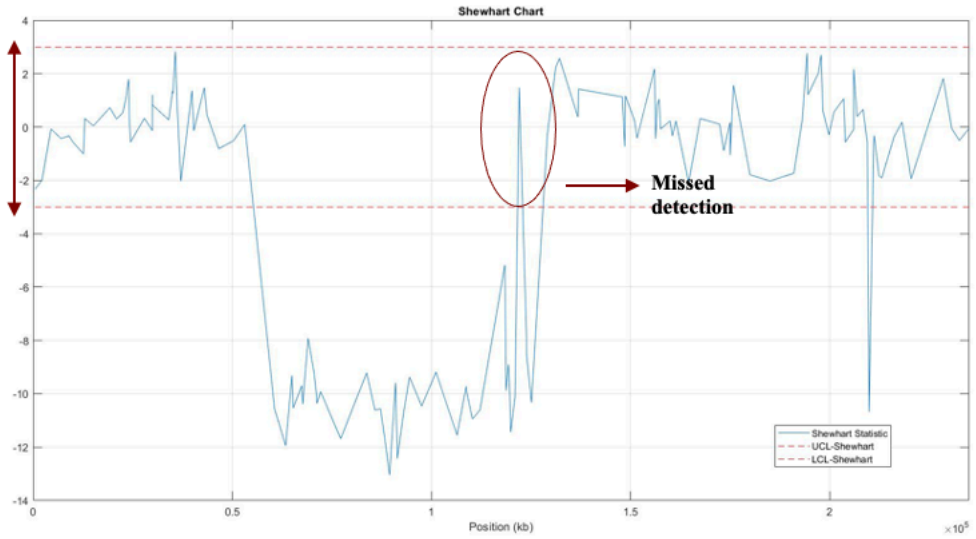

Figure 3-10:Application of fault detection on real copy number data (Shewhart chart)

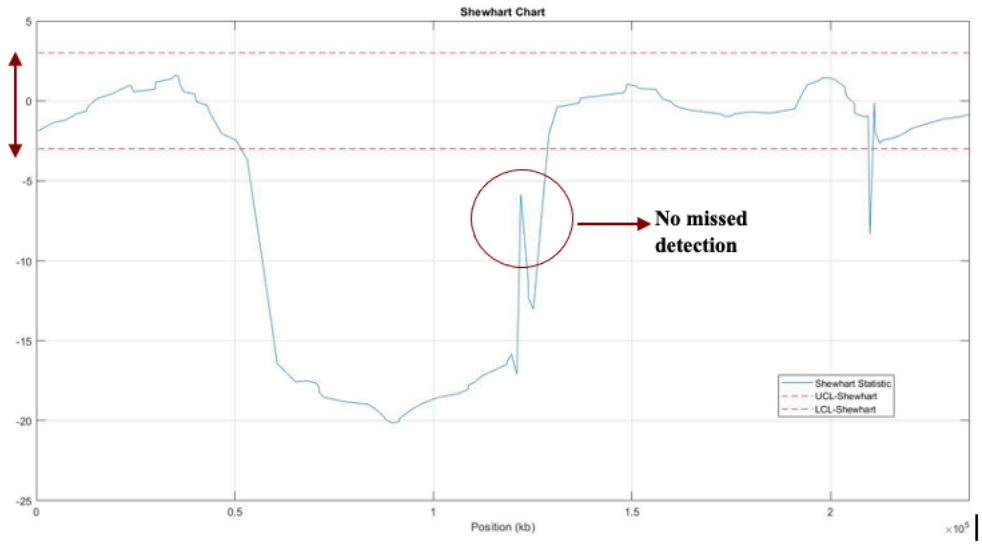

Figure 3-11:Application of fault detection on filtered copy number data (Shewhart chart)

4. CONCLUSION

This report successfully demonstrates how certain statistical filtering and fault detection techniques improve the accuracy of disease diagnosis by enhancing the accuracy of determining the locations of such aberrations. Some of these techniques include multiscale wavelet-based filtering and hypothesis testing based fault detection. The filtering techniques include Mean Filtering (MF), Exponentially Weighted Moving Average (EWMA), Standard Multiscale Filtering (SMF) and Boundary Corrected Translation Invariant filtering (BCTI). The fault detection techniques include the Shewhart chart, EWMA and Generalized Likelihood Ratio (GLR).

The performance of these techniques was illustrated using Monte Carlo simulations and through their application on real copy number data. Based on the Monte Carlo simulations, the non-linear filtering techniques performed better than the linear techniques, with BCTI performing with the least error. This is because linear filters define a frequency threshold above which all features are considered noise. This leads to important features being deleted due to their high frequency and keeping noise due to their low frequency. As for the fault detection techniques, GLR had the lowest missed detection rate at a fixed false alarm rate while Shewhart chart had the highest. This is due to the fact that Shewhart Chart does not have memory whilst the other fault detection techniques do. GLR performed the best as it is a hypothesis based technique. The application of these techniques on cancer cell lines corroborated the results obtained from Monte Carlo simulation. Applying the filtering techniques on the raw data before the fault detection techniques further enhances their performances. This shows that these techniques can be a helpful tool in disease diagnosis.

REFERENCES

[1] Y. Wang and S. Wang, "A novel stationary wavelet denoising algorithm for array-based DNA Copy Number data.," *Int. J. Bioinform. Res. Appl.*, vol. 3, no. 2, pp. 206–222, 2007, doi: 10.1504/IJBRA.2007.013603.

[2] I. Lobo, "Copy Number Variation and Genetic Disease," *Nat. Educ.*, vol. 1, no. 1, 2008, [Online]. Available: https://www.nature.com/scitable/topicpage/copy-number-variation-and-genetic-disease-911/.

[3] C. Aouiche, X. Shang, and B. Chen, "Copy number variation related disease genes," *Quant. Biol.*, vol. 6, no. 2, pp. 99–112, 2018, doi: 10.1007/s40484-018-0137-6.

[4] F. Zare, M. Dow, N. Monteleone, A. Hosny, and S. Nabavi, "An evaluation of copy number variation detection tools for cancer using whole exome sequencing data," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–13, 2017, doi: 10.1186/s12859-017-1705-x.

[5] M. . Tham and A. Parr, "Succeed at one-line validation and reconstruction of data," *Chem. Eng. Prog.*, vol. 90, no. 05, pp. 46–56, 1994.

[6] R. D. Strum and D. E. Kirk, *First Principles of Discrete Systems and Digital Signal Processing*. Addison-Wesley, 1989.

[7] M. N. Nounou and B. R. Bakshi, "On-line multiscale filtering of random and gross errors without process models," *AIChE J.*, vol. 45, no. 5, pp. 1041–1058, 1999, doi: 10.1002/aic.690450513.

[8] W. A. Shewhart, *Statistical Method from the Viewpoint of Quality Control*, 1st ed. Dover Publications, 1939.

[9] N. Johnson and F. Leone, "Cumulative Sum Control Charts: Mathematical principles applied to their construction and use. Part III," *Ind. Qual. Control*, vol. 19, no. 2, pp. 22–28, 1962.

[10] S. W. Roberts, "Control Chart Tests Based on Geometric Moving Averages," *Technometrics*, vol. 1, no. 3, pp. 239–250, 1959, doi: 10.1080/00401706.1959.10489860.

[11] M. R. Reynolds and J. Lou, "An evaluation of a GLR control chart for monitoring the process mean," *J. Qual. Technol.*, vol. 42, no. 3, pp. 287–310, 2010, doi: 10.1080/00224065.2010.11917825.

[12] M. R. R. Jr. and J. Lou, "A GLR Control Chart for Monitoring the Process Variance," *Front. Stat. Qual. Control 10*, pp. 3–17, 2012.

[13] M. R. Reynolds, J. Lou, J. Lee, and S. Wang, "The design of GLR control charts for monitoring the process mean and variance," *J. Qual. Technol.*, vol. 45, no. 1, pp. 34–60, 2013, doi: 10.1080/00224065.2013.11917914.

[14] M. Z. Sheriff, M. N. Karim, H. N. Nounou, and M. N. Nounou, "Process monitoring using PCA-based GLR methods: A comparative study," *J. Comput. Sci.*, vol. 27, pp. 227–246, 2018, doi: 10.1016/j.jocs.2018.05.013.

[15] M. Z. Sheriff and M. N. Nounou, "Enhanced performance of shewhart charts using multiscale representation," in *2016 American Control Conference (ACC)*, Jul. 2016, pp. 6923–6928, doi: 10.1109/ACC.2016.7526763.

[16] M. Z. Sheriff, "Improved Shewhard chart using multiscale representation," Texas A&M University, 2015.

[17] M. N. Nounou, H. N. Nounou, N. Meskin, A. Datta, and E. R. Dougherty, "Multiscale denoising of biological data: A comparative analysis," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 9, no. 5, pp. 1539–1544, 2012, doi: 10.1109/TCBB.2012.67.

[18] A. M. Haque, "Enhanced Monitoring Using Multiscale Exponentially Weighted Moving Average Control Charts," no. August, 2016.

[19] M. N. Nounou, H. N. Nounou, and M. Mansouri, "Model-based and model-free filtering of genomic data," *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 2, no. 3, pp. 109–121, Sep. 2013, doi: 10.1007/s13721-013-0030-1.

[20] J. Y. Keller, L. Summerer, M. Boutayeb, and M. Darouach, "Generalized likelihood ratio approach for fault detection in linear dynamic stochastic systems with unknown inputs," *Int. J. Syst. Sci.*, vol. 27, no. 12, pp. 1231–1241, 1996, doi: 10.1080/00207729608929330.

[21] M. Z. Sheriff, M. N. Karim, H. N. Nounou, and M. N. Nounou, "Process monitoring using PCA-based GLR methods: A comparative study," *J. Comput. Sci.*, vol. 27, pp. 227–246,

2018, doi: 10.1016/j.jocs.2018.05.013.

[22] F. Y. Nan, R. D. Nowak, F. Y. Nan, and R. D. Nowak, “Generalized likelihood ratio detection for fMRI using complex data,” *IEEE Trans. Med. Imaging*, vol. 18, no. 4, pp. 320–329, 1999, doi: 10.1109/42.768841.

[23] M. Z. Sheriff, C. Botre, M. Mansouri, H. Nounou, M. Nounou, and M. N. Karim, “Process Monitoring Using Data-Based Fault Detection Techniques: Comparative Studies,” in *Fault Diagnosis and Detection*, vol. 32, no. tourism, InTech, 2017, pp. 137–144.

[24] A. M. Snijders *et al.*, “Assembly of microarrays for genome-wide measurement of DNA copy number,” *Nat. Genet.*, vol. 29, no. 3, pp. 263–264, 2001, doi: 10.1038/ng754.

APPENDIX: COPY NUMBER DATA (SW837 AND MPE600)

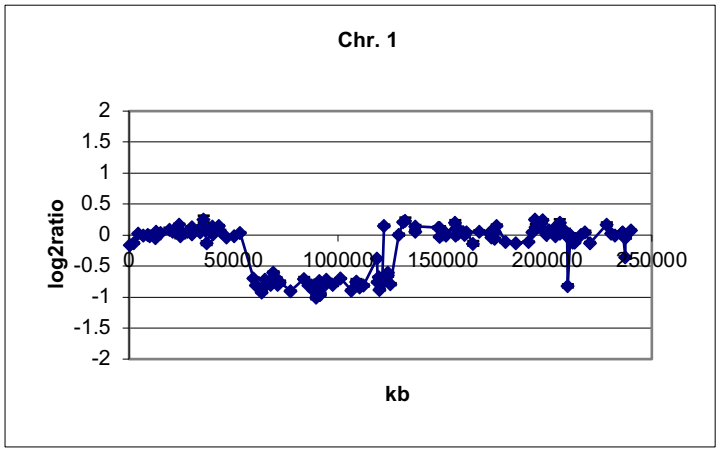

Figure A.1: Colorectal cancer cell line (SW837).

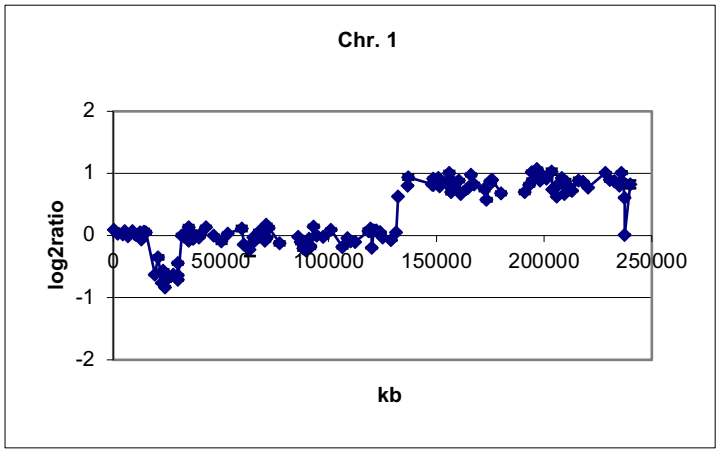

Figure A.2: Breast cancer cell line (MPE600).